MODELING SPEECH AND AUDIO CODECS REVERBERATION ARTIFACT

Yves Zango [1,2,3], Régine Le Bouquin Jeannès [2,3] and Catherine Quinquis [1]

[1] Orange Labs - Lannion, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2] INSERM, U 1099, Rennes, F-35000 France
[3] Université de Rennes 1, LTSI, Rennes, F-35000, France
{yves.zango, catherine.quinquis}@orange.com, regine.le-bouquin-jeannes@univ-rennes1.fr

ABSTRACT

Speech and sound codecs subjective assessment requires anchor signals to allow the comparison of results from different laboratories. The anchor signal presently used, the Modulated Noise Reference Unit (MNRU), is based on the hypothesis that coding technique artifact is only due to quantization noise. Earlier work showed that the impairments of speech codecs could be described by a four-dimensional event and two of these dimensions, namely "Muffled" and "Background noise" dimensions, have been already modeled. In this paper, we propose to design anchor signals for the third dimension mostly characterized by "Echo/Reverberation" attribute.

Index Terms — Speech coding, Audio coding artifacts, Modified Discrete Cosine Transform, Reverberation, Echo.

1. INTRODUCTION

Audio communications remain the most used service in telecommunication systems. Furthermore, audio and video multimedia devices, such as MP3 players, are getting more and more place on technologies markets. The proliferation of telecommunications companies and multimedia manufacturers forces them to assure the best quality of their services and devices. In a telecommunication, the phone call quality depends on the equipments involved in the transmission process. In the particular case of voice communication, the quality is highly correlated to the choice of voice and sound codecs. This choice becomes more critical, since the codecs used in the different equipments of the transmission channel are not always identical. To assess speech and sound codecs, two methods exist: subjective and objective assessment. The subjective assessment of audio codecs consists in asking a group of subjects to rate them. This kind of evaluation requires anchor signals for many reasons, for example to help the subject in his rating task or to allow the comparison of results obtained by different laboratories. Until now, the only available anchor signal is the recommendation P.810 of normalization section of ITU (International Telecommunication Union) also known as MNRU (Modulated Noise Reference Unit) [1] [2]. The improvement of electronics and signal processing encouraged the design of new codecs which implement different complex techniques, making the MNRU obsolete.
In a recent work, Etamé *et al.* [3] demonstrated that speech and sound coding techniques artifacts can be represented in a four-dimensional perceptual space. In this paper, we propose to design anchor signals for the third dimension identified as the "Echo/Reverberation" dimension.
The paper is organized as follows. First of all, we recall the four-dimensional space of speech codecs. Secondly, we present the principle of psychoacoustic model-based codecs using the MDCT (Modified Discrete Cosine Transform) transform analysis [4]. The fourth section is devoted to the description of the reverberation anchor signal design algorithm. In the fifth section, we present the results of dissimilarity and verbalization task aiming at validating the technique we used to design the anchor signals and, finally, we draw some conclusions.

2. FOUR-DIMENSIONAL SPACE

To determine the perceptual dimensions for present speech and audio codecs quality, a dissimilarity test has been already elaborated on a set of 20 tandems/codecs [3] listed in Table 1. The coding techniques implemented in each of the tandems/codecs are presented in Table 2.
Recently, we carried out a second dissimilarity test on 20 stimuli corresponding to the signals obtained by coding the original signal using the 20 tandems/codecs presented in Table 1. This test aimed at validating the anchor signals of the two first dimensions [5]. Moreover, we ran a verbalization task to label these dimensions.
This validation dissimilarity test reinforced the four-dimensional space found in [3], while the result of the verbalization task allowed qualifying these dimensions respectively as "Muffled", "Background noise", "Echo/Reverberation" and "Distorted speech" dimensions. The design and validation of the two first dimensions anchor signals have been already derived from this first study. We also observed in [5] that the "Echo/Reverberation" dimension was highlighted by the positioning of stimuli 17, 18, 19 and 20 based on MDCT technique (see Tables 1 and

2). Therefore, this dimension appears as representative of transform coding technique impairments.

Index	Description	Index	Description
1	G722.1C_24kbps_x2	11	G722_56kbps_x2
2	G722.1C_24kbps_x3	12	G722_56kbps_x3
3	G722.1_24kbps_x2	13	G729.1_14kbps_x3
4	G722.1_24kbps_x3	14	G729.1_20kbps_x3
5	G722.2_12.65kbps_x2	15	G729.1_24kbps_x2
6	G722.2_12.65kbps_x3	16	G729.1_32kbps_x3
7	G722.2_15.85kbps_x2	17	HEAAC_24kbps_x2
8	G722.2_8.85kbps_x2	18	HEAAC_32kbps_x2
9	G722_48kbps_x2	19	MP3_32kbps_x1
10	G722_48kbps_x3	20	MP3_32kbps_x2

Table 1 – Codecs/tandems under assessment (x2 and x3 mean respectively that tandem speech coding is applied two and three times to the considered codec)

Codecs	Technical characteristics
G722.1C	Modulated Lapped Transform (MLT)
G722.2	Algebraic Code Excited Linear Prediction (ACELP)
G722	Waveform codec
G729.1	Hybrid codec
HEAAC	Modified Discrete Cosine Transform (MDCT)
MP3	

Table 2 – Technical description of codecs under assessment

3. PSYCHOACOUSTIC AND TRANSFORM CODING TECHNIQUE PRINCIPLE

The MDCT-based codecs represented by stimuli 17 to 20 in our study are psychoacoustic model-based audio codecs. Figure 1 synthesizes the principle of the common architecture of these codecs.

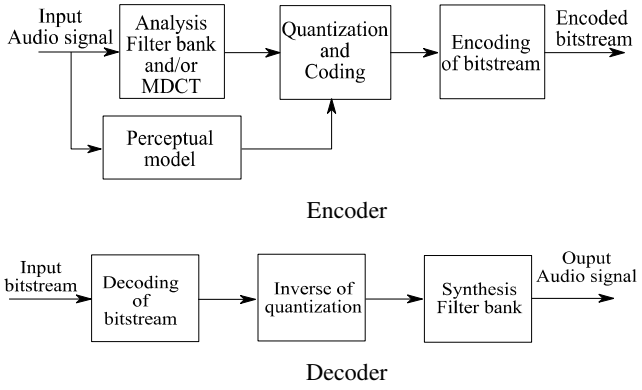

Fig. 1. Basic structure of perceptual encoder and decoder

The analysis filter bank performs a time-frequency analysis by splitting the signal into several spectral subbands to better quantize particular frequencies. In the case of MP3 [6] (stimuli 19 and 20) and HE-AAC [7] (stimuli 17 and 18) it is a polyphase filter bank. The analysis filter bank is often followed by a MDCT technique, which is the case of MP3 and HE-AAC codecs. The MDCT is useful for two main reasons. First, it allows the cancellation of time domain aliasing due to the filter bank. Secondly, it splits each subband given by the analysis filter bank to achieve a high frequency resolution.

Denoting x the signal and $x(n)$ the n^{th} sample of the signal, the corresponding k^{th} MDCT coefficient is defined by:

$$X_k = \sum_{n=0}^{2N-1} w(n)\,x(n)\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right) \quad (1)$$

where $w(n)$ is a window allowing a perfect reconstruction of the signal of $2N-$point length. This window must respect some rules (e.g. be symmetric) and preserve the energy in order to get a perfect reconstruction at the decoder side. The sine window defined by equation (2) respects these conditions and is commonly used

$$w(n) = \sin\left(\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right). \quad (2)$$

The psychoacoustic block tries to model the human audition perception in order to quantize first the most perceptible frequencies. When the MDCT is used, the psychoacoustic model guides the choice of the window to use.

The quantization and encoding steps consist in allocating dynamically bits while maintaining the quantization noise below the masking threshold determined by the psychoacoustic model. Particularly MP3 and HE-AAC use a Huffman encoding technique.

The decoder realizes the inverse operations of the encoder. The encoded bitstream is first decoded. Then, the inverse quantization is performed. Finally, the reconstructed signal is obtained by applying a synthesis filter corresponding to the analysis filter bank of the encoder.

4. ECHO-REVERBERATION ANCHOR SIGNALS

The anchor signals to be built must display impairments that can be controlled and measured easily in order to tile the perceptual space. Moreover, we are looking for a process easy to reproduce. To design the anchor signals relative to the reverberation dimension, we tried to mimic the transform coding technique principle. As seen in the previous section, we started by applying to the original signal a transform coding technique using a sine window with a 50% overlap. The quantization and Huffman encoding were skipped. Due to the bitrate limitation, all coefficients issued from the transform analysis were not quantized. The most perceived ones were quantized first respecting the psychoacoustic model, and the least perceived ones might be not quantized at all when bits are running out in the quantization step. The most perceived frequencies were likely to be the most energetic ones. Moreover, the non-quantized coefficients could correspond to different frequencies from one frame to the next one. Consequently, we proposed the following method to reproduce these phenomena.

We substituted the analysis filter bank and MDCT blocks by a STFT (Short-Time Fourier Transform). This substitution aimed at reducing the computational load. The signal was processed on a frame by frame basis with a 50% overlap using a sine window of 10 milliseconds. In order to preserve the energy amount of the original signal, in each frame, we decided to keep the k (in percent) most energetic frequency bins. The psychoacoustic model was replaced by a random loss of the least energetic frequency bins. The process was performed on a frame-by-frame basis in two main steps. As the number of the most energetic frequency bins in each frame was not constant, the choice of a fixed value of k for each frame allowed getting the same loss rate in all frames. In practice, for each frame i, we computed the number of frequency bins m_i that had a magnitude larger than the averaged magnitude in the frame. We computed the mean m of the m_i which revealed to correspond to 10% of the frequency bins in a frame. So, keeping this percentage in each frame, k was set to 10. To generate the artificial reverberation phenomenon, in each frame, we erased randomly l percent of frequency bins on the remaining ones (corresponding to 90% of all frequency bins), so that the impairment degree was controlled by the value of this parameter l.

To summarize, the function realizing the "Echo/Reverberation" artifact had as inputs the parameters k and l, the first one depending on the input signal characteristics (even if in the present work the parameter k was systematically set to 10), and the second one regulating the rate of reverberation in the output signal.

5. DISSIMILARITY AND VERBALIZATION TESTS

To assess the relevance of our anchor signals we carried out a subjective test. The test consisted of two parts: a first part was a dissimilarity test in which subjects were asked to rate the distance they perceived between pairs of stimuli. The second part was a verbalization task. During this part of the test, the listeners had to describe qualitatively the impairment they perceived on the stimuli, using a list of predefined qualifying attributes.

5.1. Dissimilarity test

5.1.1. Dissimilarity test process

To keep the duration of the test acceptable by the listeners, we limited the stimuli to one sample. This sample was taken from the phonetically balanced double sentences of France Telecom database. The original signal was a French double sentence uttered by a male speaker: "*La vanille est la reine des arômes. Fragile, il ne résiste pas à l'air glacé.*" The sentences were separated by a short silence and the total duration of the signal was 6 seconds. The stimuli were obtained by coding the original signal by the 20 tandems/codecs listed in Table 1. Since some codecs were Super-Wideband ([50-14000 Hz]) or Fullband ([20-22000 Hz]) codecs, all stimuli were downsampled at 16 kHz before conducting the test in order to avoid the influence of bandwidth on the listener's judgment.

We added to the 20 stimuli of the database four anchor signals with "Echo/Reverberation" artifact. In a first step, we generated ten anchor signals whose loss percentage l varied from 10% to 100% by step of 10% (according to the process described in section 4). Then, we retained four of these ten signals whose loss percentage was respectively 10%, 30%, 60% and 90%. In the following, we labeled them respectively stimuli 21, 22, 23 and 24.

The dissimilarity test was finally run on the 24 stimuli above, and the listeners had to compare a total of 276 $\left(C_{24}^2\right)$ pairs of stimuli to which we added the 24 null pairs. A null pair of stimuli was composed of two identical stimuli. They were introduced to test the reliability of the listeners. The rating scale was from 0 (the two stimuli of a pair are perceived as identical) to 100 (stimuli are perceived as completely different).

Twenty-one listeners were recruited for the tests. To preserve the listeners from the fatigue, the test was split in two sessions of 150 stimuli each. At the beginning of each session, the listeners had to carry out a training phase to accommodate themselves to the test rules. The training session consisted in rating 8 random pairs of stimuli among the 300 pairs. After the training session, we analyzed with each listener his score to check if the test rules were well understood. After the second session and a short pause, subjects had to carry out the verbalization task.

5.1.2. Dissimilarity test analysis

For each listener we obtained a dissimilarity matrix. In order to quantify the reliability of the listeners, we analyzed the dissimilarity score they gave to null pairs. Theoretically, this score had to be zero. For each listener, the scores of the null pairs were on the diagonal of his dissimilarity matrix. The analysis of the diagonal of the 21 dissimilarities matrices gave rise to the exclusion of only one of listeners for whom the mean of the diagonal was higher than 10. Given the remaining 20 dissimilarity matrices, we performed a three-way MultiDimensional Scaling (MDS) [8] analysis using PROXSCAL (PROXimity SCALing) algorithm [9]. The three-way MDS allowed performing a dimensionality reduction and the iterative algorithm PROXSCAL that we chose minimized the approximation error called normalized raw stress. This analysis was realized using the IBM software SPSS 19. We plot on Figure 2 the normalized raw stress curve for a number of dimensions varying from 2 to 10.

This curve shows an elbow around the fourth dimension so that we retained a four-dimensional configuration. According to this result, we conclude that the generated anchor signals did not modify the number of dimensions of

the perceptual space found in [3]. In other terms, the anchor signals did not induce supplementary artifacts leaving the perceptual space stable, which is consistent with previous studies [3].

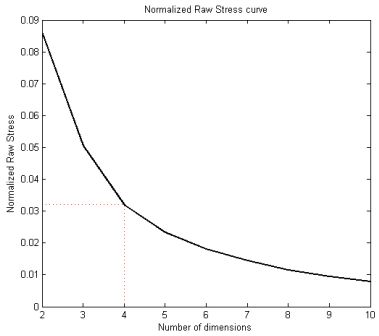

Fig. 2. The normalized raw stress curve

In order to identify the dimensions, we computed the correlation between the dimensions of the initial perceptual space [3] and the dimensions of the new space.
As shown on Table 3, dimensions 1 and 3 (Dim 1 and Dim 3) keep the same ranking (correlations equal respectively to -0.96 and 0.73). The second dimension of the initial space is correlated with the fourth dimension of the validation space.

		Initial space			
		Dim 1	Dim 2	Dim 3	Dim 4
Validation space	Dim 1	-0.96	0.19	-0.04	0.08
	Dim 2	0.07	0.11	0.21	-0.39
	Dim 3	0.09	-0.21	0.73	0.49
	Dim 4	-0.1	-0.74	-0.52	0.37

Table 3 – Correlation between the initial perceptual space and the new space

As we see on Figure 3, the MDCT codecs having the worst quality, *i.e.* stimuli 17 (the lowest bitrate of the HEE-AAC codecs) and 20 (the highest tandeming order of MP3 codecs), are located at the positive extremity of the third dimension. Moreover, the coordinates of the anchor signals (stimuli 21, 22, 23 and 24) are logically organized along this dimension.

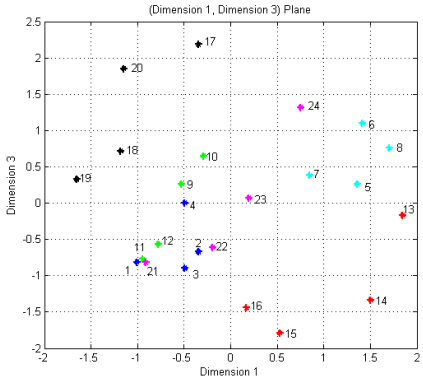

Fig. 3. Stimuli plot in (Dimension 1, Dimension 3) plane

5.2. Verbalization task

5.2.1. Verbalization task description

As the listeners' vocabulary varied, we tried to aggregate synonymous attributes to get finally four different groups presented in Table 4. During the verbalization task, the 24 stimuli were submitted to the listeners in a random order. For each stimulus, the listeners had to choose the attributes the most suited to the impairment they perceived.

Group 1	Reverberation, Echo, Metallic voice, Robot voice
Group 2	Background noise
Group 3	Muffled, Energy variation
Group 4	Distorted speech, Crackling, Modulated noise

Table 4 – Verbalization attributes

5.2.2. Verbalization task analysis

We analyzed the attributes given by the listeners to the generated anchor signals (stimuli 21 to 24). We found that these stimuli were generally characterized by the attributes belonging to the first group. Figure 4 illustrates the distribution of the different attributes used to describe the four anchor signals. It corresponds to the percentage of listeners who used a given group of attributes to describe the anchor signals. Regarding stimulus 24, the groups of attributes "Echo", "Background noise" and "Muffled" were all represented to describe this stimulus. The group of "Distorted speech" attribute was never used by the listeners, whereas the "Echo" attribute was the most contributive (61.5%). For stimulus 23, the "Echo" attribute was still mainly used (44.8%). As expected, for a limited loss of frequency bins (10% and 30%), the attributes were more uniformly distributed. As an example, for stimulus 21, the attributes "Echo", "Background noise" and "Muffled" and "Distorted speech" contributed respectively to 24%, 32%, 40%, and 4%. To conclude, stimuli 23 and 24, which were the most degraded signals, were mostly described by "Echo/Reverberation" attribute, which was not the case for the two other anchor signals.

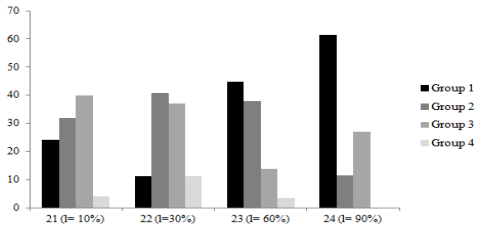

Fig. 4. Distribution of the attributes for the four anchor signals

Furthermore we studied the relative distribution of the "Echo/Reverberation" attribute for all stimuli. Figure 5 displays the occurrence of the "Echo/Reverberation" attribute for each codec compared to the number of times this attribute was used by all listeners to qualify all stimuli. As presented on this figure, this attribute was mostly used to describe three "families" of stimuli. The first one was

composed of stimuli 23 and 24 (the most "reverberant" anchor signals), the second one corresponded to the MDCT codecs (stimuli 17 to 20), and the last one was the MLT codecs family (stimuli 1 to 4). The MDCT and MLT codecs are both based on transform coding technique. Therefore, these results tend to prove that the "Echo/Reverberation" attribute can be ascribed to artifact due to "transform coding". However, we must acknowledge that the listeners also qualified stimulus 18 as "Distorted speech" (see Table 5). Concerning the G.729.1 codecs (stimuli 13 to 16), they are hybrid codecs and include a MDCT technique, which can partly explain why stimuli 14 and 16 sounded "reverberant" (see Table 5). Nevertheless, referring to Table 5, we note that "Muffled" was the most dominating attribute for this family of stimuli, except for stimulus 16 for which the most often used attributes were "Background noise" and "Distorted speech".

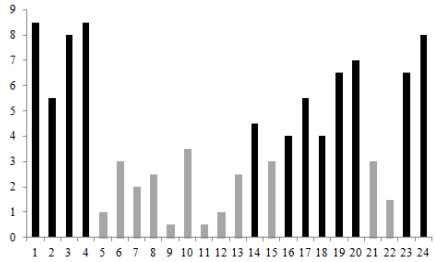

Fig. 5. The Echo/Reverberation attribute representation for each stimulus relatively to the others

Stimuli	Group 1	Group 2	Group 3	Group 4
1	65.4	11.5	15.4	7.7
2	45.8	4.2	41.7	8.3
3	64	8	24	4
4	68	4	28	0
5	8	8	72	12
6	25	8.3	66.7	0
7	19	9.5	66.7	4.8
8	18.5	0	59.3	22.2
9	3.8	26.9	15.4	53.8
10	20.6	17.6	11.8	50
11	4.2	41.7	12.5	41.7
12	8	32	8	52
13	17.9	10.7	57.1	14.3
14	29	12.9	45.2	12.9
15	19.4	16.1	41.9	22.6
16	28.6	32.1	7.1	32.1
17	37.9	6.9	20.7	34.5
18	33.3	4.2	16.7	45.8
19	52	4	8	36
20	53.8	7.7	11.5	26.9
21	24	32	40	4
22	11.1	40.7	37	11.1
23	44.8	37.9	13.8	3.4
24	61.5	11.5	26.9	0

Table 5 – Distribution in percentage of the different groups of attributes for each stimulus

6. CONCLUSION

The obsolescence of MNRU encouraged us to create a new system of anchor signals. Contrary to the mono-dimensional character of speech and audio codecs quality considered by the MNRU system, the proposed study considered this quality as multidimensional. Previous works already suggested that the perceptual space of present audio codecs might be restricted to four dimensions. Considering the two first dimensions of this perceptual space already modeled [3], we presented in this contribution a technique aiming at designing anchor signals for the third dimension. This dimension was representative of "Echo/Reverberation" artifact due to transform coding technique. To model it, we simulated the MDCT coding technique using STFT techniques and the degradation was controlled by a loss of frequency bins. A dissimilarity test and a verbalization task allowed us to validate the relevance of our approach. On the one hand, the most degraded anchor signals were the most qualified as "reverberated" signals by the listeners. On the other hand, the anchor signals were well-ordered on the third dimension. The last dimension was mostly qualified as "Distorted speech" and our ongoing research is devoted to the design of corresponding anchor signals, the final goal being to validate a new reference system to replace MNRU.

7. REFERENCES

[1] H.B. Law, R.A. Seymour, "A reference distortion system using modulated noise," *IEEE*, pp. 484-485, November 1962.

[2] ITU-T Recommendation P.810, "Modulated Noise Reference Unit (MNRU)," International Telecommunications Union, 02/96.

[3] T. Etame, R. Le Bouquin Jeannès, C. Quinquis, L. Gros and G. Faucon, "Towards a new reference impairment system in the subjective evaluation of speech codecs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, Issue 5, July 2011.

[4] Princen J, Bradley A, Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34, No.5, Oct 1986, pp. 1153-1161.

[5] Y. Zango, R. Le Bouquin Jeannès, N. Costet and C. Quinquis, "Identification of perceptive dimensions of speech and audio codecs subjective quality," *EUSIPCO 2011,* Barcelona, August 29 - September 2, 2011.

[6] T. Sakamoto, M. Taruki, T. Hase, "Fast MPEG-audio layer III algorithm for a 32-bit MCU," *IEEE Transactions on Consumer Electronics*, vo1. 45, pp.986-993, no. 3, August, 1999.

[7] J. Herre and J. M. Dietz, "MPEG-4 high-efficiency AAC coding," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137-142, May 2008.

[8] I. Borg, P J.F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, 2nd Edition, Springer, New York, 2005.

[9] J. Commandeur and W.J. Heiser, "Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Symmetric Data Matrices," *In Research Report RR-93-04*, DDT, Leiden University, 1993.